AWARD NUMBER: W81XWH-13-1-0020

TITLE: Health-Terrain: Visualizing Large Scale Health Data

PRINCIPAL INVESTIGATOR: Ph.D. Fang, Shiaofen

CONTRACTING ORGANIZATION: Indiana University
Indianapolis, IN 46202

REPORT DATE: December 2015

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
December 2015	Final	7Mar2013 - 30Sep2015

4. TITLE AND SUBTITLE
Health-Terrain: Visualizing Large Scale Health Data

5a. CONTRACT NUMBER
W81XWH-13-1-0020

5b. GRANT NUMBER
12108017

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Shiaofen Fang, Mathew Palakal, Yuni Xia, Shaun J. Grannis, Jennifer L. Williams

E-Mail: craigjen@regenstrief.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Indiana University
Indianapolis, IN 46202

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The promise of the benefits of fully integrated electronic health care systems can only be realized if the quality of emerging large medical databases can be characterized and the meaning of the data understood. For this purpose, the effective visualization of large and complex health data for timely decision making is critical. Our long-term goal is to improve the usability of emerging large scale clinical data sets by developing effective and efficient open-source systems for health data analytics and visualization tools for clinicians, healthcare professionals, administrators, and patients. The objective of this application is to develop a prototype system to test the effectiveness of this approach on a large scale health care database that is currently available at Regenstrief Institute. We have reached this objective with the following specific accomplishments:

- Built a relational database as the representation of a health concept space, extracted from the NCD dataset.
- Natural Language Processing techniques were carried out to process 325791 clinical notes to extract new terms including diseases, symptoms, and mental and risky behaviors.
- Data mining techniques were applied to extract associations between terms in the concept space, and to discover new cluster terms.
- Designed and implemented a suite of novel visualization algorithms that allows the users to interactively explore the data based on the user selected terms and filters.
- Designed and implemented a web based graphical user interface for the prototype system.
- Designed and tested an evaluation procedure for health data visualization system.

This visualization framework offers a real time and web-based solution for the effective use of large scale military electronic health record systems by allowing system level integration of the human´s visual capabilities into the overall health data based decision making system. The visual representation of concept space provides a method to compress large, heterogeneous, and historical patient and public health data into a single, intuitive and comprehensive visualization. The new spatiotemporal visualization techniques developed here are novel and particularly suited for large public health datasets that involve geographical and population wide information.

15. SUBJECT TERMS

Information visualization; Visual analytics; Public health data, Notifiable condition detector; Text mining; Data mining.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
Unclassified	Unclassified	Unclassified	Unclassified	30	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

INTRODUCTION

The goal of this project is to develop novel visualization techniques and tools for large and complex health care data to facilitate timely decision-making and trend/pattern detection. A prototype system will be developed to test the effectiveness of this approach on a large-scale health care database that is currently available at Regenstrief Institute. More specifically, we want to develop a public health use case leveraging a Notifiable Condition Detector (NCD) dataset that contains reportable disease conditions that are transmitted to Indiana public health authorities (over 800,000 reports). Clinicians and public health stakeholders seek to uncover informative trends contained within the growing population-based datasets. To support knowledge discovery, we first extract meaningful terms and their associations and attributes from the raw data by applying data mining and text mining algorithms to construct a concept space. A browser-based user interface is developed to enable interactive online data exploration. A suite of visualization algorithms and techniques are developed and implemented within the prototype system.

KEYWORDS
Information visualization, Visual analytics, Public health data, Notifiable condition detector, Text mining, Data mining

OVERALL PROJECT SUMMARY

The project had four primary goals, all of which were accomplished.

I. Concept space definition
II. Algorithm design
III. System design and implementation
IV. System Prototyping and Usability Evaluation.

Concept Space Definition

The "concept space" represents a uniform layer of clinical observations and their associations and serves as a platform for users to explore the data using visualization and analysis methods. The concept terms are derived from data mining and text-mining processes applied to the use case datasets. For this project we focus on a population health use case that leverages an automated Notifiable Condition Detector (NCD). The NCD dataset contains 833,710 notifiable cases spanning more than 10 years from among 439,547 unique patients [1]. An additional dataset linked to the original NCD patient's data was extracted from the Indiana Network for Patient Care (INPC) health information exchange containing 325,791 unstructured clinical discharge summaries, laboratory reports, and patient histories [2]. Disease concepts were extracted from the NCD dataset. Text mining algorithms were then applied to additional linked text dataset (unstructured clinical summaries) to construct ontologies for different concept types, including Disease, Symptom, Mental behavior, and Risky Behavior. An association-mining algorithm was applied to the combined terms to generate an association graph among all the concepts terms. The resulting concept space, along with the processed NCD data, is represented in a data model designed to support our specific ontology.

Data Model Design

Considering the visualization-specific requirements, we designed a three-layer data model (Figure 1) to store the NCD and supporting text dataset. The first layer contains base tables for the entities included in our ontology: patient, disease, location, and other terms. The table for these additional terms has four subcategories: mental behavior, risky behavior, medication and symptom. The second layer contains associations between the primary patient entity and additional three supporting entities. The third layer contains indirect associations between disease, term and location and was constructed using data mining techniques. Designs for the specific supporting schema and classes of associations were informed by the data mining results and the data elements necessary to support each specific visualization. Further, to avoid costly database scans during visualization execution the schema also includes pre-computed aggregate data necessary to support the specific visualizations. Pre-computed aggregate data include joint statistics such as entity association frequencies, e.g., the number of instances of "disease X" associated with "location Y".

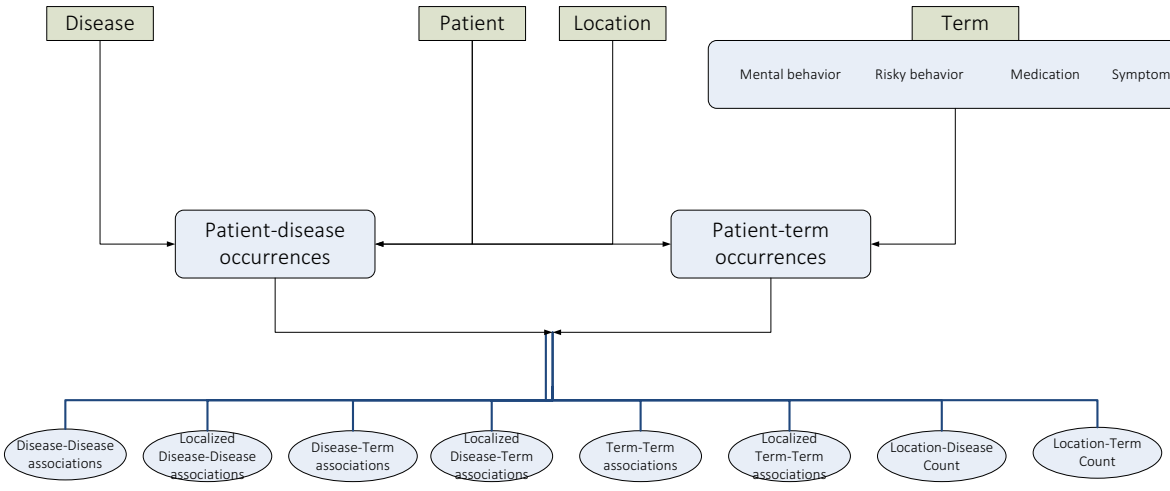


Figure.1 Database model

Data Cleansing

To preserve patient confidentiality we created a randomly assigned, pseudonymized patient identifier, (called "PseudoID") linking records within and among the NCD and INPC datasets, but conveys no identifiable traits. In rare instances a PseudoID may falsely match more than one patient. To avoid this error, we used three additional fields to verify that records sharing the same PseudoID represent the same patient. The three additional fields are gender, race, and date of birth. If two records have the same PseudoID but disagree on non-null genders with values of 'M' or 'F', then the records are treated as separate patients. The race validation rule functions similarly to gender validation. For date of birth validation, we standardize date of birth format as "yyyy-mm-dd" and apply the longest common subsequence string comparator. If the ratio between the length of the longest common subsequence over the length of yyyy-mm-dd format is less than a certain threshold, date of birth validation fails. Records are determined to represent the same patient only if all three additional fields pass validation.

After cleansing, the database contained 439,547 patients, 1,976 diseases, 3,756 locations and 3,851 terms (711 symptoms, 93 risky behaviors, 200 mental behaviors and 2847 medications). The second layer of the database contains 1,302,173 disease occurrences and 1,215,659 term occurrences. At least 90,376 patients are associating with at least one term non-disease. All of these patients have a least one disease. The number of patients having more than one disease is 114,820, which is later used for association mining. At the third layer, the database contains 577,888 global associations between two different diseases, 1,958,227 global associations between two different terms and 1,032,864 global associations between a disease and a term.

We removed duplicate public health case reports, which were defined as record having the same patient, the same date, and the same notifiable condition. We subsequently identified the most common reported conditions. The condition "Lead Exposure" was found among 256,823 patients. However, lead poisoning is not common in practice. "Lead Exposure" has the highest occurrence because Indiana's reporting law requires that all laboratories performing blood lead tests report the results of those tests, whether normal or abnormal. Therefore, even when the test result is in the normal range, the test was reported. It leads to a high number of records on "Lead Exposure" in the data, while most of the report has negative results. Additional frequently reported conditions included: 1) Staphylococcus Methicillin-Resistant Aureus (MRSA), 2) HIV, 3) Chlamydia Infection, 4) Hepatitis B, 5) Hepatitis C, 6) Gonorrhea, 7) Chickenpox, 8) Measles, 9) Hepatitis A, 10) Enterococcus Vancomycin-Resistant (VRE) 11) Trichomoniasis 12) Syphilis. Figure 2 shows the number of occurrences of the most common diseases.

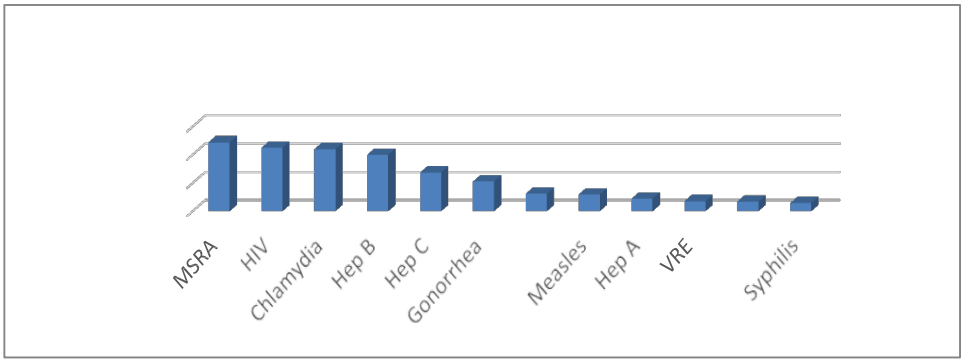


Figure 2: Most Common Diseases in the NCD dataset

We analyze the disease distribution across races. Here we compare the difference between the two largest races: white and black. The result is shown in Figure 3, with the black bar representing the occurrence percentage of each disease among black patients and the blue bar representing the occurrence percentage of disease among white patients. It shows that among black patients, CHLAMYDIA INFECTION and GONORRHEA are the most common conditions in the NCD data. TRICHOMONIASIS and SYPHILIS are also more common in black patients than in white patients. Among white patients, the most common condition is STAPHYLOCOCCUS METHICILLIN-RESISTANT aureus (MSRA).

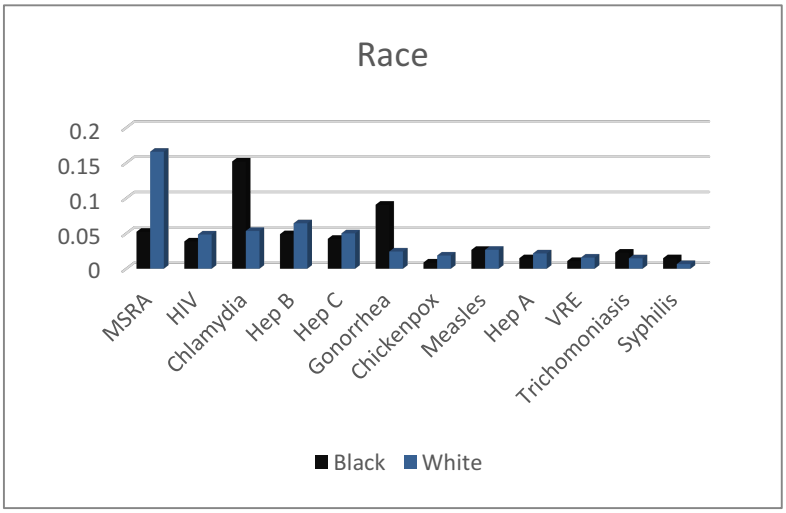


Figure 3: Diseases Distribution Across Race

Query efficiency

We test the database efficiency by three sets of common samples queries designed by visualization and health science experts. The first query set is about geographical distribution of one or a combination of diseases. The second query set retrieves strong associated diseases to a given disease. The third query set finds common diseases occurring at a given range of age. Table 1 summarizes the performance of three types of queries and suggests that the further optimization will be required for effective user interactions during interactive visualization.

Query set	Example	Involved tables	Runtime
1	Geographical (at city level) distribution of chlamydia	location, diseases, patient-disease occurrences	12s
2	List the diseases associating with chlamydia	diseases, associations	0.5s
3	What are the most common diseases for patient age from 20 to 40	diseases, patients, patient-disease occurrences	16s

Table 1: 3 query set for testing database

Algorithm Design

There are 3 types of algorithms we have developed: (1) Text Mining algorithms to extract concept terms from clinical notes; (2) Data mining algorithms, such as association mining and clustering; and (3) visualization algorithms for various visualization tools.

Text Mining

The NCD receives clinical data that includes the diagnoses, laboratory studies, and transcriptions from hospitals, national labs and local ancillary service organizations. But much of the information pertaining the patients' condition is available in the clinical reports. Mining these reports can provide a bigger picture of various other conditions that the patient experienced during his/her treatment. This information can provide valuable insights on the patients' socioeconomic condition, behavior risk factors, environmental factors and genetic information (family history). Natural Language Processing (NLP) provides a means to augment the NCD data analytics with the information discovered from these clinical reports.

Text Mining Process

NLP techniques were carried out to process 325791 clinical notes that contain patient discharge summaries, laboratory reports, patient history, etc. Although these records are de-identified due to which the patient specific information are

absent, a pseudo-patient Id has been provided to help process the reports. Basic processing of the reports was performed for converting the clinical notes from XML format to simple text format and sentence splitting. Advanced level NLP was applied in the form of named entity recognition (NER) for extracting diseases, symptoms, mental behavior, risky behavior and medication information from the reports. This was done with the help of UMLS [3] database which is a repository of clinical and health related terms. Once the entities were extracted using NER, negation analysis was applied using NEGEX algorithm [4] to remove negated terms. Figure 4 shows the process that was used in extracting this vital information from the reports.

The advantage of using UMLS is that all variations of clinical terms get captured that provide a large set of terms available for further analysis. For example, clinical notes that indicate "Hepatitis" contains terms like "Hepatitis", "Hepatitis B", "Hep", "Hep B" etc. The large number of terms extracted contains different occurrences of the same diseases, symptoms, etc. We apply stemming and grouping algorithms to reduce the total number of terms. The identified terms are stored in different data tables and joined using the pseudo-patient Id.

Comorbidity Analysis

Once the data tables are constructed, we perform deeper analysis to compute the comorbid conditions of the diseases. For this, we use the *tf-idf* (term frequency – inverse document frequency) vector space model [5] to identify the significantly co-occurring diseases. The *tf-idf* model is considered to be an effective text mining model that provides the importance of a term/word to a document in a collection of documents. This model uses the concept of relevance and co-occurrence of terms. Equation (1) gives the relevance of a term *j* w.r.t. a document *i*,

$$w_{ij} = t_{ij} * \lg(\frac{N}{N_j}) \quad (1)$$

where w_{ij} = relevance of term *j* in the patient record *i*; t_{ij} = term frequency of term *j* in in the patient record *i*; N_j = frequency of records for term *j*; *N* = total number of records (*N*=325791).

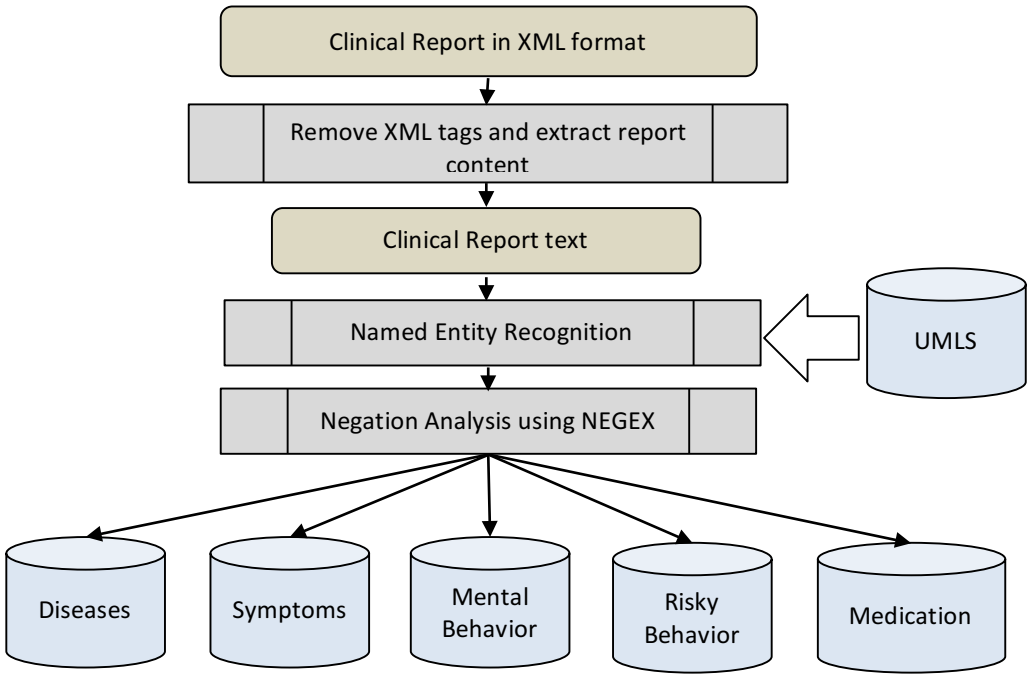


Figure 4. NLP steps applied on Clinical Reports

A particular term is more relevant *w.r.t.* a record if it appears more frequently in the record and appears in fewer numbers of records in the total records set. An association weight/score is attached with every association between a pair of terms [5]. This is given by A_{jk}

$$A_{jk} = \sum_{i=1}^{N} t_{ij} * \lg(\frac{N}{N_j}) * t_{ik} * \lg(\frac{N}{N_k}) \quad (2)$$

This is essentially a product of the relevance of each of the pair of terms over the entire records set *N*. The association score is 0 if the terms do not co-occur in any of the *N* records. Associations with non-zero scores are considered to be associated to the term.

After applying basic level processing on the reports, the clinical content from the reports was subjected to NER. UMLS was used for NER to identify the diseases, symptoms, mental behavior, risky behavior and medication terms from the 325791 reports. The total number of terms extracted for each category is given in Table 2. Figure 5 shows the most commonly occurring diseases with the number of reports in which they were found.

The top 10 diseases were analyzed using the *tf-idf* model to identify comorbidity of the diseases across the 325791 reports. To achieve this, we compute the pair-wise significance of each disease with all the corresponding conditions, (i.e., the symptoms, mental behavior, risky behavior and medications) using Equation 2. Table 3 shows the top 10 diseases and the corresponding conditions.

Term Type	Number of terms extracted using NLP
Diseases	7988
Symptoms	10803
Mental Behavior	712
Risky Behavior	244
Medications	5721

Table 2: Total terms identified by NLP

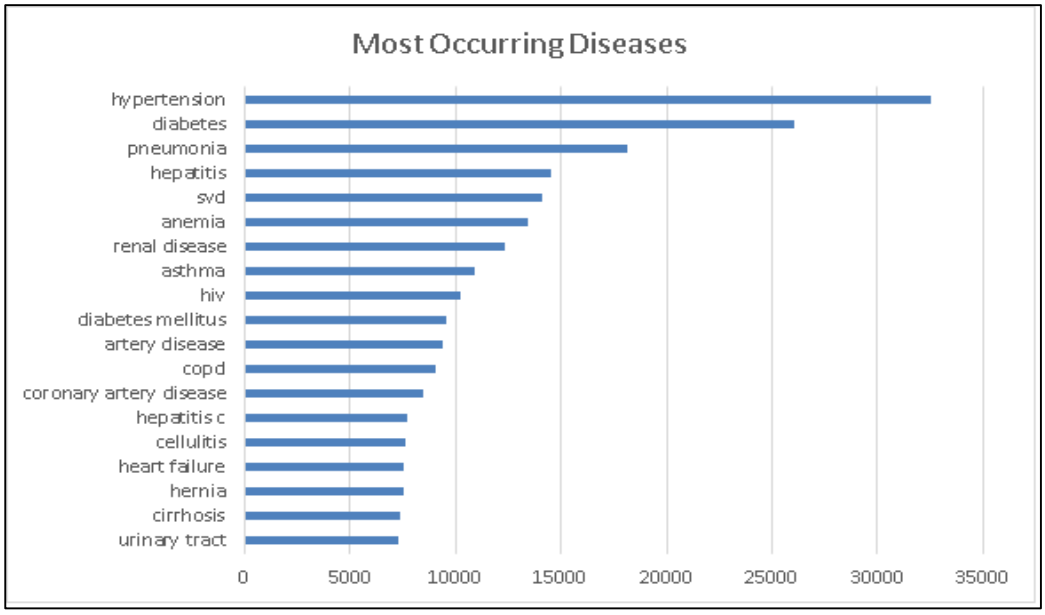


Figure 5: Most commonly occurring diseases and corresponding number of reports

Disease Name	Diseases	Symptoms	Mental Behavior	Risky Behavior	Medications
hypertension	diabetes, renal disease, pulmonary hypertension, artery disease,	chest pain, nausea, vomiting, dyspnea, abdominal pain, weakness,	abuse, depression, dementia, anxiety, altered mental status, drug use,	smoking, tobacco use, compliance, impression, drinking, lying,	insulin, hepatitis, tobacco, oxygen, glucose, lasix,
diabetes	diabetes mellitus, hypertension, artery disease, renal disease,	nausea, vomiting, chest pain, abdominal pain, diarrhea,	abuse, depression, altered mental status, drug use,	smoking, compliance, tobacco use, impression, drinking,	insulin, glucose, tobacco, hepatitis, humulin,
pneumonia	lower lobe pneumonia, aspiration pneumonia, copd,	shortness of breath, chest pain, dyspnea, chills, vomiting,	abuse, dementia, aggressive, confusion,	smoking, impression, drinking, tobacco use, compliance,	oxygen, avelox, albuterol, prednisone, levaquin,
hepatitis	hepatitis c, hepatitis b, cirrhosis, liver disease, encephalopathy,	nausea, abdominal pain, vomiting, diarrhea, chills,	abuse, dependence, confusion, drug use, opiate, depression,	smoking, drinking, tobacco use, illicit drug use,	hepatitis, hepatitis b, prograf, lactulose, ammonia, antibody,
svd	gbs, pcc, ofc, strep, hep, external genitalia,	prn pain, constipation, cramping, headache,	abuse, drug use, depression, substance, substance abuse,	smokes, illicit drug use, smoking, tobacco use,	micronor, vitamin, antibody, ibuprofen, stool softener,
anemia	renal failure, diabetes, hypertension, renal disease, hepatitis, heart failure,	nausea, abdominal pain, vomiting, chest pain, fatigue, weakness,	abuse, depression, anxiety, dementia, confusion, altered mental status,	smoking, drinking, impression, tobacco use, compliance,	iron, vitamin, hepatitis, coumadin, oxygen, prednisone,
renal disease	end-stage renal disease, end stage renal disease, diabetes, hypertension, artery disease,	nausea, vomiting, chest pain, abdominal pain, chills, shortness of breath,	altered mental status, abuse, confusion, dementia, depression, confused,	smoking, compliance, impression, tobacco use, illicit drug use, drinking,	calcium, insulin, glucose, coumadin, hepatitis, bicarbonate,
asthma	pneumonia, diabetes, copd, hypertension, airway disease,	wheezing, shortness of breath, wheezes, coughing, dyspnea,	abuse, depression, mdi, anxiety, drug use, aggressive,	smoking, impression, drinking, tobacco use, crying,	albuterol, prednisone, medrol, oxygen, atrovent, advair,
hiv	aids, pneumonia, hepatitis, infectious disease, herpes, meningitis,	nausea, vomiting, diarrhea, abdominal pain, headache, weakness,	abuse, depression, schizophrenia, drug use, dementia, dependence,	compliance, smoking, drinking, impression, lying, tobacco use	hepatitis, bactrim, vitamin, cocaine, acetaminophen, hepatitis b,
diabetes mellitus	diabetes, hypertension, artery disease, renal disease,	vomiting, nausea, chest pain, abdominal pain, diarrhea,	abuse, depression, altered mental status, dementia,	smoking, tobacco use, compliance, illicit drug use,	insulin, glucose, humulin, tobacco, hepatitis,

Table 3**:** Comorbid conditions with top 10 most occurring diseases

Symptom	Occurance	Mental Behavior	Occurance	Risky Behavior	Occurance
vomiting	11	abuse	11	smoking	11
abdominal pain	10	depression	11	tobacco use	10
chest pain	10	anxiety	10	compliance	10
nausea	10	drug use	10	impression	10
weakness	9	altered mental status	9	drinking	9
diarrhea	8	confusion	9	illicit drug use	6
dyspnea	8	dementia	8	lying	6
shortness of breath	8	confused	5	crying	2
chills	7	drug abuse	5	grunting	1
headache	4	aggressive	4	marijuana	1
constipation	2	dependence	3	sobriety	1

Table 4**:** Most comorbid behaviors for the top 10 diseases

We also analyzed the most common conditions that occurred with these diseases (Table 4). It was interesting to find that well known behaviors such as “smoking”, “depression” and “tobacco use” were amongst the commonly occurring conditions.

Data Mining

Association mining

In layer 3, we compute and store two types of associations. The first type is the conditional probability, or rule confidence, between two entities. Given two different entities i and j, the rule confidence between *i* and *j* is computed as

$$\text{Rule\_confidence}(i, j) = \frac{|i \wedge j|}{|i|}$$

in which $|i \wedge j|$ is the number of patients showing both entities *i* and *j* and | *i* | is the number of patients showing entity *i*. The second type of association shows the happen-before relationship between entities *i* and *j*, and is computed as the probability that entity *i* detection time is before entity *j* detection time

$$\text{Happen\_before}(i, j) = \frac{|i \text{ before } j|}{|i \wedge j|}$$

in which $|i \text{ before } j|$ is the number of showing *i* before showing *j*. We only compute localized association when then number of patients in the location is above 1000.
The database contains significant associations which are not widely reported in literature, such as Antidiarrheal treatment and runny nose symptom (confidence: 0.73), sclera and Tylenol treatment (confidence 0.70), posturing and Motrin treatment (confidence: 0.81), etc. Table 5 shows part of the association rules in tabular format. The premise and conclusion of the rule is shown in the table, with the quality measures of each rule including the support, confidence, Laplace, Gain, p-s, lift and Conviction. We are working with domain experts on evaluating the association rules and tuning the parameters to produce optimum result.

No.	Premises	Conclusion	Support	Confiden...	LaPlace	Gain	p-s	Lift	Convic...
5	SYPHILIS	HUMAN IMMUNODEFICIENCY VIRUS	0.010	0.286	0.976	-0.060	0.007	3.468	1.285
6	HEPATITIS A	HEPATITIS C	0.028	0.294	0.939	-0.162	0.017	2.626	1.258
7	HEPATITIS A	HEPATITIS B	0.032	0.332	0.942	-0.159	0.009	1.423	1.148
8	MUMPS	CHICKENPOX	0.010	0.348	0.981	-0.050	0.008	4.377	1.413
9	MEASLES	CHICKENPOX	0.033	0.368	0.948	-0.145	0.026	4.628	1.457
10	CHICKENPOX	HEPATITIS B	0.029	0.370	0.954	-0.130	0.011	1.589	1.218
11	MEASLES	HEPATITIS B	0.036	0.403	0.951	-0.142	0.015	1.726	1.284
12	HEPATITIS C	HEPATITIS B	0.045	0.404	0.940	-0.179	0.019	1.732	1.287
13	CHICKENPOX	MEASLES	0.033	0.412	0.957	-0.126	0.026	4.628	1.548
14	ENTEROCOCCUS VANCOMYCIN-RESISTANT	STAPHYLOCOCCUS METHICILLIN-RESISTANT	0.019	0.442	0.977	-0.067	0.014	3.847	1.586
15	MYCOBACTERIUM NON-TB	AFB UNDETERMINED	0.011	0.450	0.987	-0.038	0.011	31.475	1.793
16	TRICHOMONIASIS	CHLAMYDIA INFECTION	0.020	0.493	0.981	-0.060	0.010	2.098	1.509
17	MUMPS	HEPATITIS B	0.015	0.497	0.985	-0.045	0.008	2.129	1.523
18	MUMPS	MEASLES	0.016	0.539	0.987	-0.044	0.014	6.065	1.978
19	CHLAMYDIA INFECTION	GONORRHEA	0.142	0.604	0.925	-0.328	0.094	2.932	2.007
20	GONORRHEA	CHLAMYDIA INFECTION	0.142	0.689	0.947	-0.270	0.094	2.932	2.460
21	AFB UNDETERMINED	MYCOBACTERIUM NON-TB	0.011	0.764	0.997	-0.018	0.011	31.475	4.143
22	TRACHOMA	CHLAMYDIA INFECTION	0.014	0.881	0.998	-0.018	0.010	3.749	6.436

Table 5: Association rules among diseases

Clustering analysis
We developed a co-clustering algorithm to cluster both diseases and text-mining terms to discover potential combinations of both diseases and terminologies, which could be disease subtypes or imply new biomedical patterns. The algorithm iteratively and partially [6] allocates the diseases or terms into clusters based on the rule confidence attributes. Let *K* be the number of clusters. To reallocate the diseases given the clustering allocation of terms, we select the *k* giving the maximum affinity score (*as* score) of disease *i* on cluster *k*. The *as* score is computed as

$$as(i,k) = \sum_{\forall j \in C_k} \left(p(i \mid j) - \overline{p(l \mid j)}_{l} \right)$$

in which C_k denotes the cluster *k*, *j* is the index of the term and $\overline{p(l \mid j)}_{l}$ is the mean of the associations given term *j*. $\overline{p(l \mid j)}_{l}$ is the repulse factor to prevent the case when all diseases and terms falls in one cluster. The process to reallocate the terms is similar to the diseases allocation process.

The algorithm can be executed in parallel by using a master-assistant computational model to improve efficiency. When reallocating diseases, the master routine sends the term-cluster allocations to all assistants and assigns the disease subsets for each assistant to reallocates. The assistants send the disease allocation results for the master routine for later use in terms allocation. We terminate the iterative allocation steps until the number of diseases/terms adopting new cluster is small and the clusters become stable.

We found several cluster containing close relationships between diseases and terminologies, such as {Biliary Sludge, HFA, Macrocytosis, Paroxysmal, Pseudogout, back discomfort, betimol, hesitancy, Intron A}, {Gastric Polyps, Kidney failure, antral, benefix, benicar}, {Duodenal Ulcer, Helicobacter Pylori, Malabsorpition, amylase, antimetics} and {appetite lost, immunoglobulin, retrovir} , etc. Some clusters highly correspond to specific diseases or medical processes. For example, appetite lost, immunoglobulin, retrovir are HIV related symptoms. Meanwhile, some clusters contain diseases and terms associated with several medical processes. For example, we found a cluster including gastrointestinal terms (Duodenal ulcer, Helicobacter pylori, Malabsorption, acyclovir, amylase and antiemetics), additive behavior (drinking, lortab andmarijuana) and cancer (methotrexate, vincristine and zofran). This cluster may suggest negative

impact of addictive behavior toward digestive system. The appearance of cancer drugs in this cluster could raise a research question about the impact of additive behavior toward the metabolism process, which will further affect the cancer drug efficiency.

Sequential pattern mining

We construct the sequence of disease/term occurrence based on rule_confidence and happen_before association. Only associations with rule_confidence and happen_before association greater than certain threshold and covering at least 50 patients are included to construct the sequence and visualization. Due to the limited number of text records showing the test date, we only applied sequential pattern mining on disease association.

We found 105 disease-associations satisfying all 3 criteria about rule_confidence, begin_before_end and coverage to construct the frequent sequential disease patterns. We found 3 groups of sequences in the NCD data. The first group contains only one sequence Hyperplenism Annemia. The second group contains 5 diseases: Fibrosis Pulmonary, Staphylococcus Methicillin-resistant, Biliary Stricture, Cycsticercosis and Meconium Ileus, in which Fibrosis pulmonary and Meconium ileus stay at the triggering position. This disease group may raise additional research questions since these diseases occur at different organs. The last group is marked by Hepatitis A and 56 other diseases staying at the triggering position of Hepatitis A.

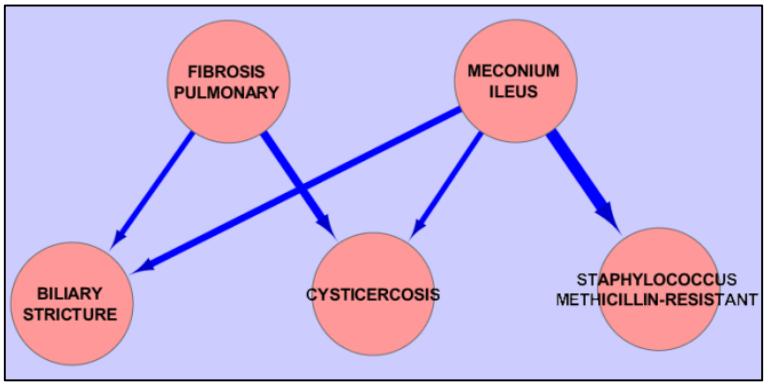


Figure 6. Fibrosis/Meconium-ileus sequence

Visualization algorithms

We have developed a suite of visualization algorithms for the interactive visual exploration of the health data represented in the concept space database.

Association graph

The opening visualization is an associative graph of the diseases and other terms from association mining. Association map is a graph visualization of the association relationships among the diseases and other terms in the concept space. It can serve as a platform supporting interactive selection of concepts to dynamically visualize data using a variety of tools in the visualization system. To draw an association graph, a spring-embedded algorithm [7] is used to layout the graph nodes by optimizing the following energy function:

$$E_s = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{2} k(d(i,j) - s(i,j))^2$$

Where *d(i,j)* is the 2D Euclidean distance of two nodes, and *s(i,j)* is a similarity metric of two nodes representing the heuristic of the layout. Edge thickness indicates the strength of association, and node size can reflects the number of other nodes to which a given node has a significant association, or the total occurrence of a term (e.g. disease) in the dataset. Nodes can be selected, and the graph will be quickly redrawn to only show other nodes which have significant association to the selected nodes.

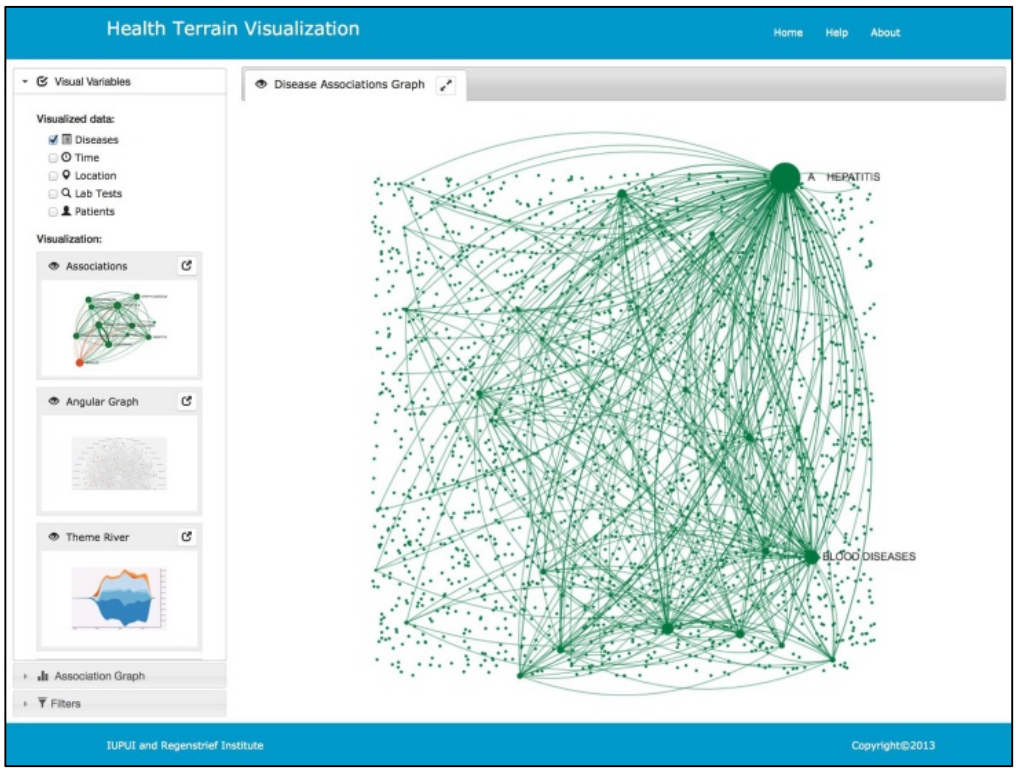


Figure 7. Disease association map and the web interface of the HealthTerrain system

Theme River

Theme river view [8] shows the aggregate trend for the terms (e.g. diseases and symptoms) selected by the user for a given time period. Each term (a theme) is visually represented as a river stream, and implemented as a filled curve plot along the horizontal time axis, with y-axis representing the occurrence of the term. Multiple themes are stacked together vertically for side-by-side comparison of the streams over time, as well as the possible interactions.

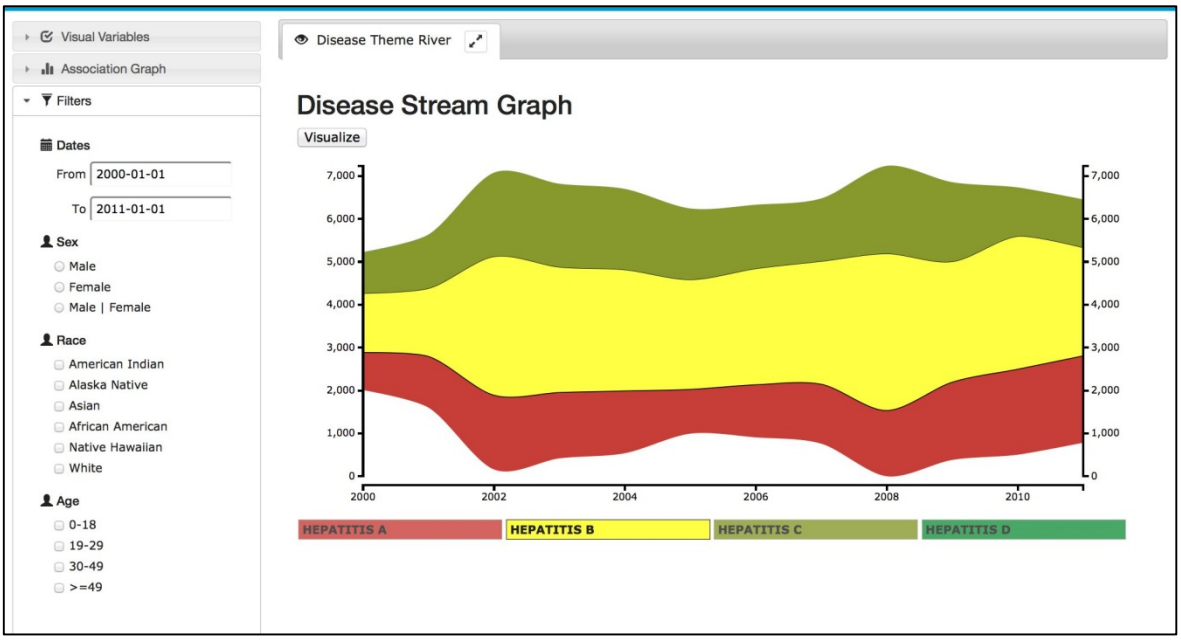


Figure 8. The Theme River visualization for Hepatitis A, B, C and D

Ring graph

In order to view more detailed patient level data, we developed a new patient visualization method called Ring Graph. In Ring Graph, each patient is modeled as a point in a radial coordinate system. The radial space is subdivided into multiple rings, each of which represents one visualization term that was selected from the association map. These terms are typical disease names, but can also be other associated terms such as symptoms and risky behaviors. The circumference of this radial space represents the time-axis. Thus, time is encoded as the radial angle of the points (patients). Ring Graph shows the distribution of patient-level data over a time-attribute space. One significant attribute, for example "age", will be represented as radius. Other attributes of the patients, such as race and gender, are represented as color and shape of the dots.

Occurrences of the same patient associated with multiple terms (e.g. diagnosed with multiple diseases) are connected with curves across the graph. A connecting curve will be highlighted when there is mouse over on the patient or the curve. Details of a patient record can also be shown by mouse over. To avoid clutter, the connecting curves are drawn with adjustable semi-transparent lines. Lowering the transparency can reveal more clearly the associations between terms. Figure 9 show an example of the Ring Graph for Chlamydia Infection, Gonorrhea, and Syphilis over a time period.

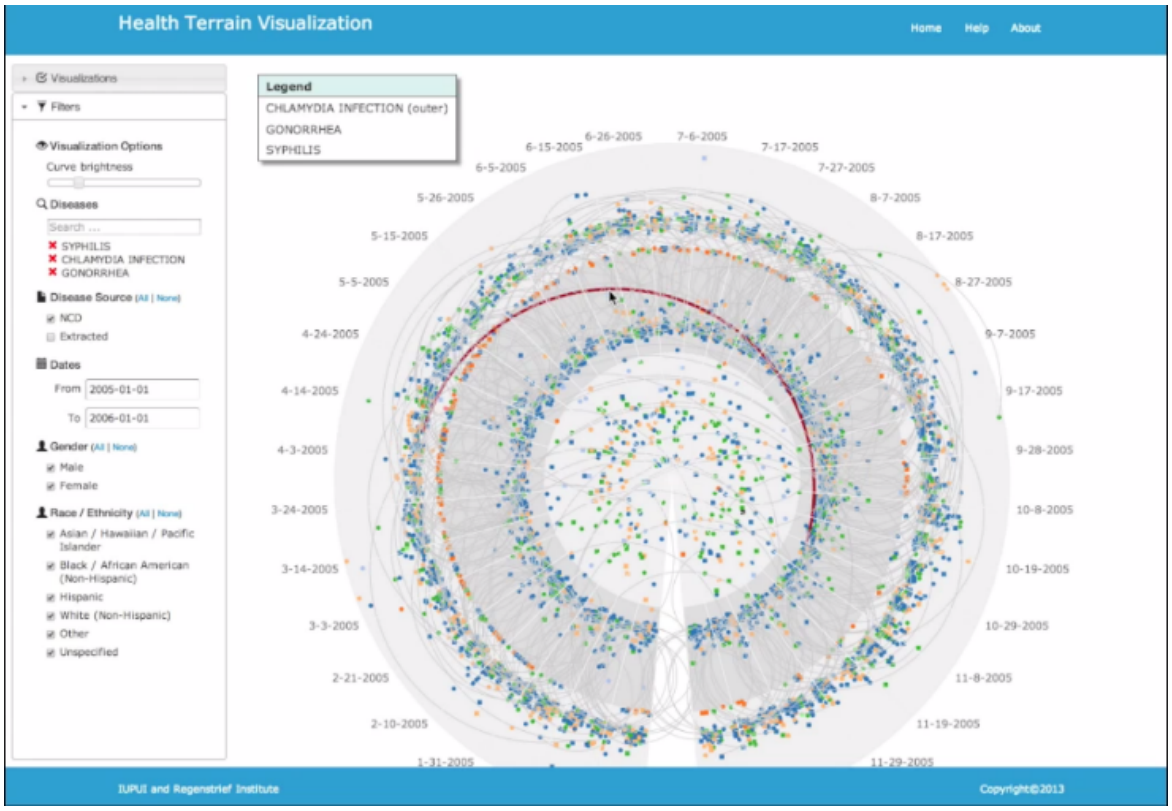


Figure 9. A Ring Graph for Chlamydia Infection, Gonorrhea, and Syphilis

Spiral Theme Plot

Ring Plot, however, does not provide a good overall trend and comparisons of different diseases over time, as typically shown in a Theme River plot. The time axis is also only limited to one circle, which cannot represent periodical patterns. We developed a new time-series visualization method called Spiral Theme Plot by integrating ThemeRiver [8] and spiral pattern [9] to plot patients as points in stacked spiral rings. Time is represented as a spiral base curve. Diseases (or any other term) are represented as stacked themes along a spiral base curve. Patients are plotted within the regions of the themes as points with proper visual attributes. One significant attribute, for example "age", will be represented as radius. Other attributes of the patients, such as race and gender, are represented as color and shape of the dots. Spiral Theme Plot allows multiple years of patients data be plotted periodically such that seasonal patterns or abnormal patterns for seasonal diseases can be easily detected. For patients with multiple hospital visits at different times for the

same or different conditions, curves are drawn to connect these multiple occurrences by the same patient. The base spiral curve is:

$$\begin{cases} x = r(\theta)\sin\theta \\ y = r(\theta)\cos\theta \end{cases}$$

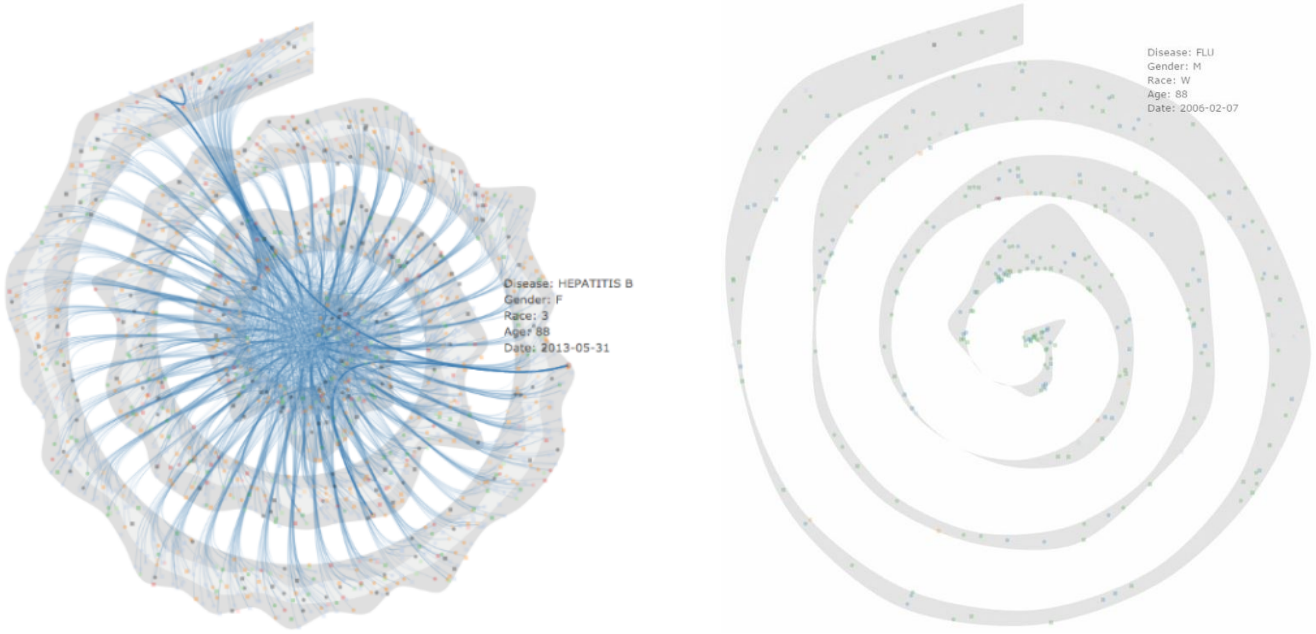
where r(θ) is a monotonic continuous radius function of angle θ. When r(θ) is a linear function r(θ)=a+bθ, the gap between the spirals is a constant 2πb, which can be estimated based on the maximum cumulative width of the themes (Fig. 10).

Figure. 10: Spiral Theme Plot for Hepatitis A, B, and C over four years.

When plotting patient data within each theme, the width of the theme at a particular angle is determined by the total occurrence of the disease at that particular time. The boundary curve of each theme can then be interpolated by spline curves. This interpolation is done by splitting the time axis into a fixed number of segments. The maximum width of each segment is used as an interpolation point. This leads to a discrete set of interpolation points from which the spline curve can be generated as the boundary curve of the theme.

When plotting a point for each patient, the width of the theme needs to be computed first in order to determine the proper radius of the point. Although this width information can theoretically be computed from the spline representations, we found that it is more efficient to simply check the color values along the normal direction of the spiral curve to estimate the width of a theme at each angle.

Lines are drawn between points representing multiple occurrences of the same patient. Such lines sometimes can become very dense leading to a cluttered image. We implemented an edge bundling strategy to bundle these connecting lines for each pre-defined time interval (Fig. 11(a)). Figure 11(b) show a periodical (seasonal) pattern of Flu over 4 years.



(a) (b)

Figure. 11: (a) Spiral Theme Plot with bundled links; (b) Seasonal pattern of Flu.

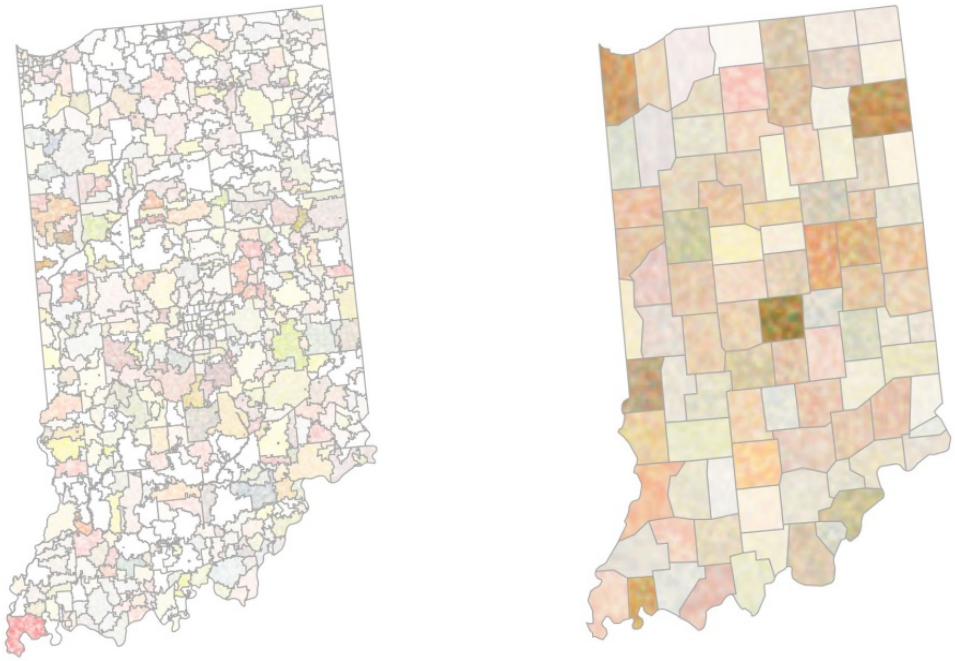
Texturization

In texturization, texture images are constructed to represent the overall data trends and distributions in different geospatial regions. Once the textures are generated, we will first visualize them on a 2D geographic map as a heatmap image, and then map them to terrain surfaces. There are two different types of textures that will be generated here (1) noise pattern texture for the representation of multiple attributes; and (2) offset contour texture for the time-varying data representation.

Noise Texture

We aim to represent multiple attributes for each geographic region using color coded texture patterns so that the users can easily perceive the representations of different attributes, not only within one region, but also its overall geospatial distributions across many regions in a geographic area (e.g. a state).

We first construct noise patterns to create a random variation in color intensity, similar to the approach in [10]. Different color hues will be used to represent different types of attributes, for example the occurrences of different diseases. A turbulence function [11] will be used to generate the noise patterns of different frequencies (sizes of the sub-regions of the noise pattern). These multi-scale patterns may be applied to different scales of geographic areas (e.g. counties vs zip-codes). Since the noise pattern involves the mixing and blending of different color hues, we choose to use an RYB color model instead of RGB model, as proposed in [10], since RYB color model provides more intuitive representation of the weights of different colors after blending. Figure 12 shows two examples of the heatmap views of three diseases, Diabetes. Hepatitis B, and Chlamydia, over the Indiana state map.



(a) (b)

Figure 12. Heatmap views of noise textures over the Indiana state map: (a) county based; (b) zip-code based

Offset Contouring

Offset contouring is designed to represent attribute changes over time within a geographic region. It can also be used to represent multiple attributes. Similar to the Noise Pattern approach, we first construct a texture image using offset contour curves to form shape-preserving sub-regions, and then use varying color shades or hues to fill the sub-regions to represent the change of attribute values over time, or to simply fill the sub-regions with different color values to represent multiple attributes. The offset contours are generated by offsetting the boundary curve toward the interior of the region, creating multiple offset boundary curves (Figure 13).

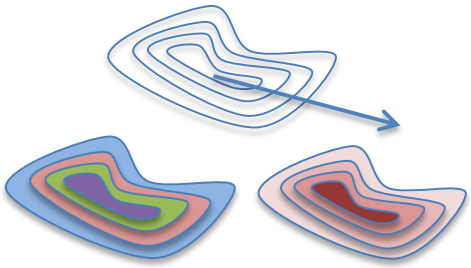
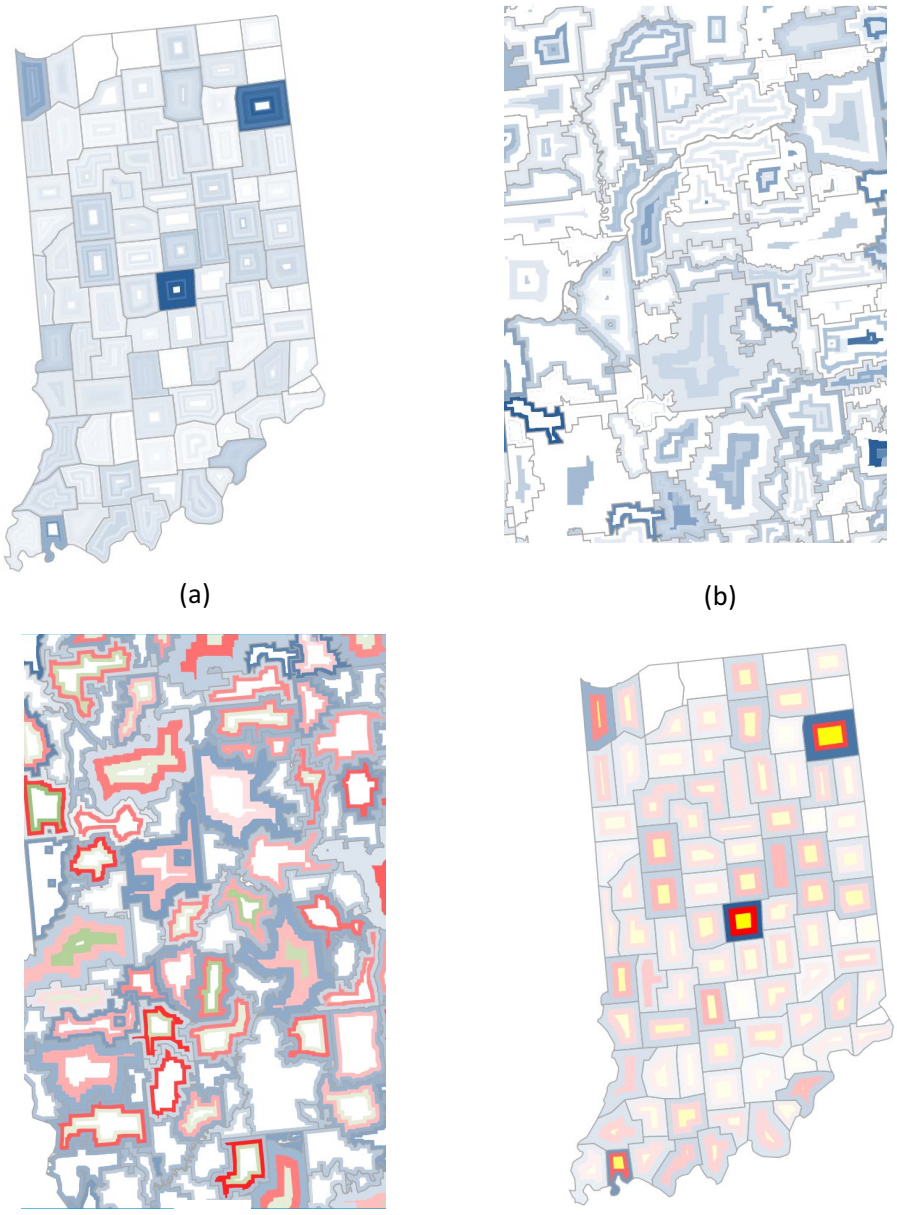


Figure 13: Offset contouring, with multi-attribute coloring and time-series coloring

There are several offset curve algorithms available in curve/surface modeling. But since in our application, the offset curves do not need to be very accurate, we opt to use a simple image erosion algorithm [12] directly on the 2D image of the map to generate the offset contours. Figure 3b and 3c shows the color-filled sub-regions after offset contouring. In time-series data visualization, the time line can be divided into multiple time intervals and represented by the offset contours. Varying shades of a color hue can be used to represent the attribute changes (e.g. occurrence of a disease) over time. This approach, however, has two limitations. First, when the boundary shape of a region is highly concave, the image erosion technique sometimes does not generate clean offset contours. This usually can be corrected using a geometric offset curve algorithm such as the one in [13]. A second limitation of this approach is that it requires a certain amount of spatial area to layout the contours and color patterns. In public health data, however, these attributes are typically defined on geographic areas, which provide a perfect platform for texturization. Figure 14 shows a few examples of the heatmap views of offset contouring over the Indiana state map.



(a) (b)

(c) (d)

Figure 14. Heatmap views of offset contouring over the Indiana state map: (a) County based time-series data; (b) Zip-code based time-series data; (c) County based multi-diseases data; (d) Zip-code based multi-diseases data.

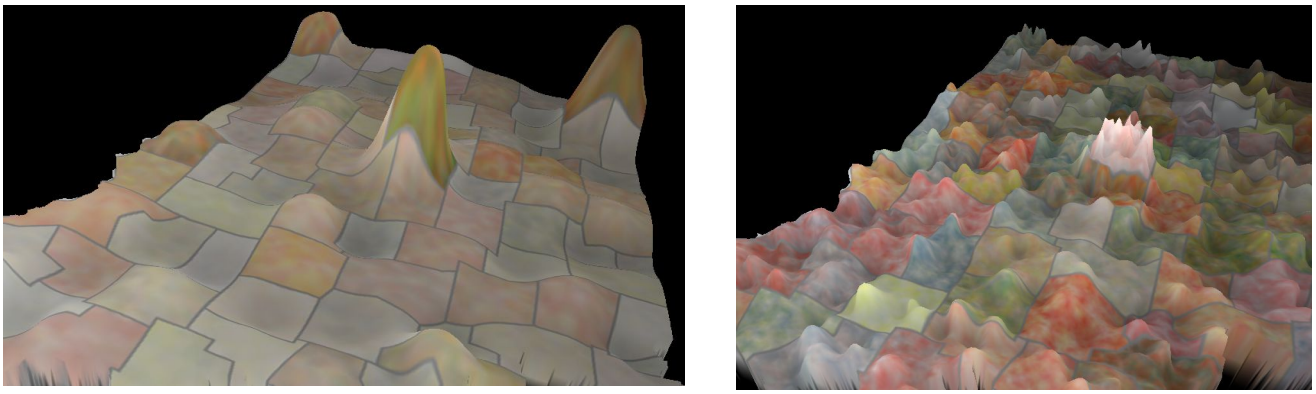
Texturized Terrain Surface

The heatmap views are effective in conveying the relative distributions of multiple attributes in different regions of a geographic area. The distribution of the total attribute values, however, becomes more difficult to perceive as the information has been disbursed by the texture patterns. This problem can be resolved by mapping the texture pattern onto a 3D terrain surface using the total attribute values as a height field.

A 3D surface can be constructed on top of a geographical region (e.g. the map of Indiana State). Typically, data are aggregated to individual geographical regions, such as counties and zip-codes, to form a height field. The height value can be, for example, the sum of multiple attribute values in a region, or the total occurrence of an attribute over the given time period for time-series data. To construct the surface, 3D scattered interpolation technique is applied so that every pixel point within the geographical boundary will have an interpolated height value. In our implementation, a Shepard interpolation method is applied:

$$d = \sum_{i=0}^{n-1} (1/r_i)^2 \cdot d_i \Big/ \sum_{i=0}^{n-1} (1/r_i)^2$$

where d is the height of an arbitrary point P within the geographical boundary, d_i are the known heights (attributes) at the known points C_i (*e.g*. center points of zip codes or counties), and r_i are the distances between P and C_i. A 2D image of the geographical map is used to limit the surface within the geographical border. This technique is implemented as a variation of our previous work on GeneTerrain [14]. Some examples are shown in Figures 15-17.



(a) (b)

Figure 15. Terrain views of a multi-disease visualization over the Indiana state map. (a) County based textures and interpolation; (b) County textures and zip-code based interpolation.

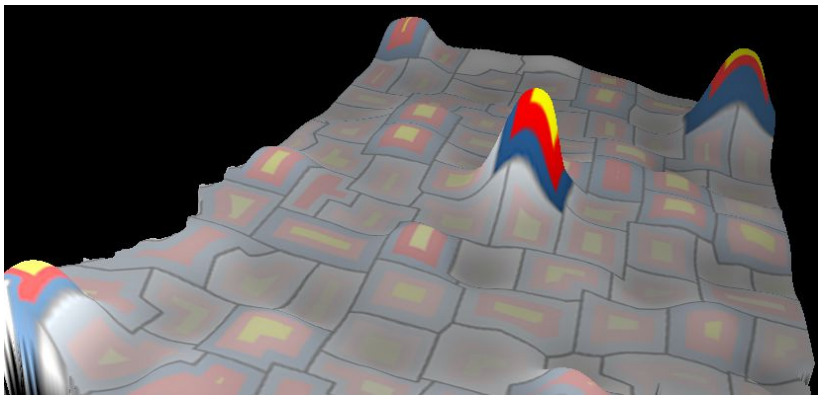
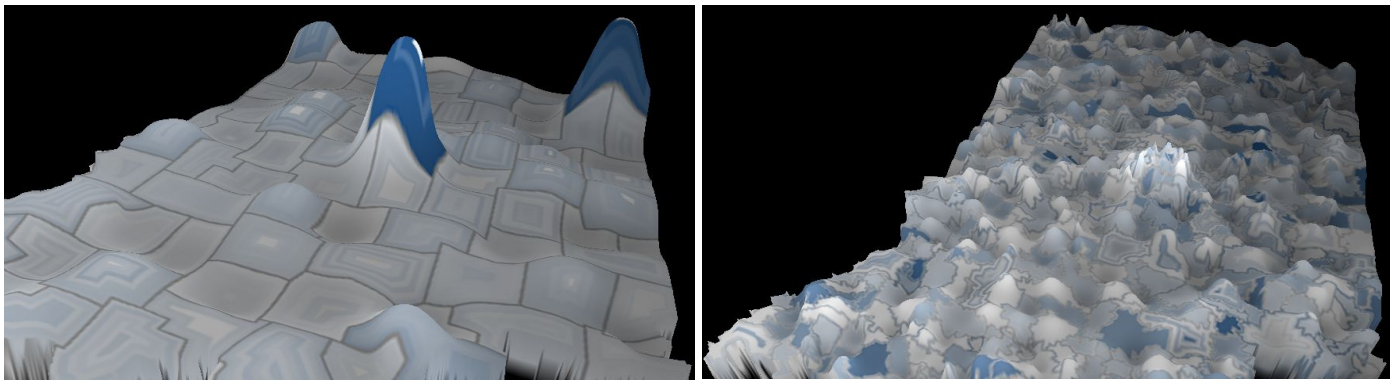


Figure 16. A terrain view of a county-based multi-diseases visualization



(a) (b)

Figure 17. Terrain views of a time-series data over the Indiana state map. (a) County based; (b) zip-code based.

System Design and Implementation

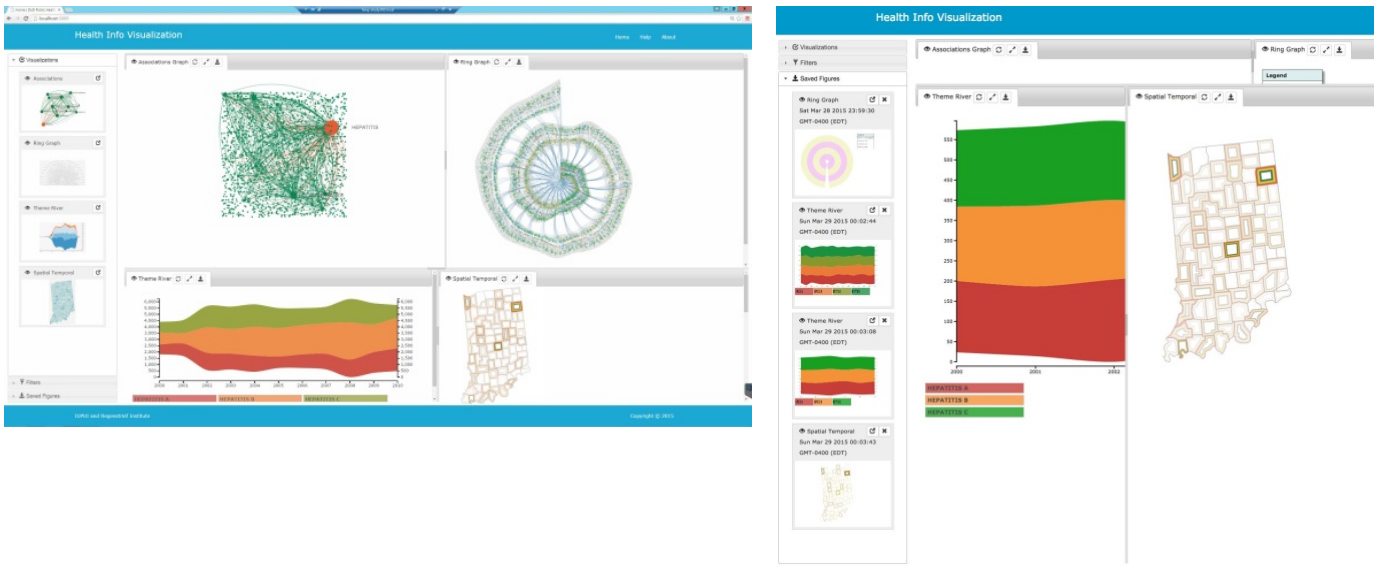
We designed and implemented the initial structure and framework of HealthTerrain visualization. Initial plans had been to develop an installed executable application written in C++ and utilizing OpenGL for interactive visualizations. However, after some initial research and experimentation in the capabilities of modern web browsers (Google Chrome, Mozilla Firefox and Apple Safari) for 2D and 3D graphics, we came to the conclusion that WebGL in an HTML5 canvas would provide sufficient technical and graphical capabilities we need while appealing to a much broader potential user base with an established and maturing set of user experience patterns. Once focused on the web, we settled on an architecture pattern based primarily on the Ruby on Rails (RoR) framework for delivering web applications with AJAX services and a classic Model-View-Controller architecture. Ruby and Rails were picked as our server-side language and framework of choice for their elegant syntax, vibrant open source community, and ease of use.

The application itself is 3-part:

1. A MySQL relational database containing the results of offline text mining and statistical analysis research on the health data set provided to us by our partners at Regenstrief Institute.
2. A server-side RoR application for querying, modeling and manipulating data in the relational database.
3. An HTML/CSS/Javascript web GUI.

The user interface is a modern web GUI utilizing a combination of form submission and RESTful service calls to query and retrieve data in various data delivery formats such as Extensible Markup Language (XML) and JavaScript Object Notation (JSON). Interactivity is a primary goal as we seek to both visualize our data and provide opportunities for novel visual exploration and analysis.

The visualizations themselves utilize HTML, CSS, SVG, and WebGL technologies with a number of open-source Javascript libraries such as sigma.js, d3.js, jquery.js and three.js for drawing, displaying and interacting with the data and graphics. Figure 16 shows a screen shot of this interface that includes multiple visualization methods in a split window interface so that the visualization of the same dataset can be compared and analyzed.



(a) (b)

Figure 15. (a) The web interface with split windows; (b) System interface with saved working windows.

System Prototyping and Usability Evaluation

After developing the data visualization framework, we imported de-identified communicable disease data and incorporated four operational visualizations including a network association graph, a ring graph, theme river graph, and 2D/3D cholorpleth (heatmap) spatiotemporal graphs. To perform a usability evaluation of this framework we recruited interviewees who represented potential end-users and visualization consumers, including public health epidemiologists with expertise in notifiable disease surveillance and syndromic surveillance; Indiana University faculty from the school of

Public health; biomedical informaticians with public health informatics expertise from the Regenstrief Institute; clinical practitioners; and program managers with advanced training in public health management.

For our usability evaluation we adapted the National Institute of Standards and Technology (2007) definition of usability for our participants as the "effectiveness, efficiency, and satisfaction with which intended users can achieve their tasks and the intended context of product use." Using an unstructured qualitative interview process, we explored dimensions of effectiveness (accuracy in completing tasks), efficiency (perceived time and effort in accomplishment of tasks), and satisfaction (subjective response to the application).

Prior to reviewing each of the four visualizations in successive order, the interviewees were oriented to the following dimensions of the application: 1) the overall screen layout and structure; 2) the ways in which users could navigate within a screen; 3) the ways in which users could navigate to other screens; 4) the ways in which users could navigate to the home screen; the ways in which users would move from field to field; and 5) a description of key commonly used buttons, icons, and links.

After presenting each visualization, the interviewees were asked to comment on the perceived dimensions of effectiveness, efficiency, and general satisfaction. Where necessary, exemplar leading questions were prepared to stimulate discussion, and included: "comment on your perceived satisfaction with the time required to interact with this visualization"; "how satisfied would you be with the perceived effort to interact with this visualization?"; "how confident are you that you could use this visualization to support your daily work flows on a routine basis?"; and "how quickly do you think most users would learn to perform the functions needed for this visualization?" The interviewees' responses were synthesized, stratified by each visualization and are summarized below:

Association Network Graph

As a general theme, the interviewees felt that including in-line guidance or pop-up descriptions (e.g., using mouse-overs) for each visualization parameter would provide end-users with valuable information to guide their use the tool. For example, the purpose of the association "threshold" parameter used in the association graph to create edges was unclear, and interviewees sought further definition. Interviewees noted that the visualization loading time, while less than 10 seconds, could be improved to enhance overall user satisfaction. The meaning of the colors of the edges in the graph was unclear, and interviewees felt they should be more clearly defined in the application. Public health stakeholders expressed the clear value of being able to quickly identify associations among multiple diseases, and they were pleased with the ability to filter out extraneous nodes and create sub-networks for strongly associated diseases. The interviewees felt that edges in the graph should contain metrics characterizing the strength of the association between nodes (disease).

Ring Graph

The interviewees described this visualization as being particularly complex and exhibiting high information density; some felt that the density obfuscated important information and were concerned that individual cases may be overlooked. The interviewees required substantial introduction to the graph prior to expressing recognition of the value of the visualization. Several commented that the extended (90-120 second) loading time was sub optimal, and hindered overall satisfaction, usability, and efficiency. While the dimensionalities of disease, age, gender, race, and time were generally perceived to be useful, the interviewees suggested that allowing those dimensions to be configurable would improve the utility of this visualization. One epidemiologist interviewee noted that their team likely would not use this visualization to identify disease outbreaks, but would instead use this visualization after an outbreak has been detected through other means in order to explore the relationships and characteristics of individuals within an outbreak in order to identify potential risk factors and target interventions. Another suggested that the circular format could be confusing and may obfuscate data; it was suggested that the graph be transformed into a linear format to potentially improve interpretability. One interviewee noted, "This graph has the potential to make me think about things that I wouldn't otherwise, and that has value to me."

Theme River Graph

Interviewees generally expressed that the theme river visualization provided a consumable, informative high-level comparison communicable disease incidence over time. Multiple interviewees indicated they would prefer case counts to begin at a common baseline on the y-axis; the variable heights and irregular sides of the theme river graph

were felt to hinder interpretability. A consistent linear y-axis baseline of zero was felt to potentially enhance year-to-year comparisons over the default theme River visualization.

Spatiotemporal Graph

The three-dimensional version of this visualization was perceived to be more informative than the two-dimensional version. Commenters noted that the two-dimensional color variations within counties were challenging to interpret; the varying color intensity combined with varying band widths for each disease confused the interviewees. Some noted that continuous variation in color intensity may be less interpretable than dividing the range of disease incidence into a discrete set of ranges. Interviewees stated that presenting disease incidence as a three-dimensional height substantially improved interpretability and understanding of the data. There is wide variation in disease rates among counties (a small number of counties contain significant portions of overall disease); this variation obfuscates details in lower prevalence regions. Consequently the interviewees suggested that an additional feature enabling nonlinear scaling to highlight details in lower prevalence counties would be useful. They further suggested that presenting these data as incident rates (new cases per total population in the county) versus absolute counts (new cases) could improve interpretability and overall satisfaction. Epidemiologist interviewees requested extended functionality to visualize the highest prevalence diseases in each county.

General observations

Due to the data privacy policy provisions of the institutional review board research process, we used obfuscated de-identified clinical data for the usability assessment. The interviewees noted that further assessment of the usability of these different visualization tools would be enhanced by reviewing fully identified data rather than the de-identified obfuscated data currently used for research and development purposes.

KEY RESEARCH ACCOMPLISHMENTS

1) Designed and implemented a MySQL relational database, as a representation of the concept space, which is derived from the NCD dataset by data mining and text mining algorithms.
2) Natural Language Processing techniques were carried out to process 325791 clinical notes to extract new terms including diseases, symptoms, and mental and risky behaviors.
3) Data mining techniques were applied to extract associations between terms in the concept space, and to discover new cluster terms.
4) Designed and implemented a suite of interactive visualization algorithms that allows the users to interactively explore the data based on the user selected terms and filters. These include the association map, the theme river graph, the ring graphs, the spiral theme plot, and the texturization based spatiotemporal visualization. The ring graph, spiral theme plot and the texturization based spatiotemporal visualization are novel techniques that has never been developed by others.
5) Designed and implemented a web based graphical user interface for the prototype system, and successfully integrated the programming interfaces between the user interface, visualization, and the database.
6) Designed and tested an evaluation procedure for health data visualization system.

CONCLUSION

We have made significant accomplishments in this project, including: (1) created a concept space definition, which represents a schema tailored to support diverse visualizations and provides a uniform ontology that allows the system to be leveraged for many types of health care datasets through individually designed text and data mining procedures; (2) designed and implemented a suite of novel visualization algorithm, as well as data and text mining analytics techniques; (3) developed a prototype visualization system for the exploration of large-scale, real-world health data; and (4) Designed and tested an evaluation procedure for health data visualization systems. These components are integrated in a generalizable browser-based graphical interface, which enables flexible and free-form data exploration and hypothesis discovery. The system has received favorable initial feedback from users, and we believe it has potential as an open source tool to support health data visualization tasks.

PUBLICATIONS, ABSTRACTS, AND PRESENTATIONS

Publications/Abstracts

Shiaofen Fang, Mathew Palakal, Yuni Xia, Sam Bloomquist, Thanh Minh Nguyen, Anand Krishnan, Shenghui Jiang, Weizhi Li, Jeremy Keiper, and haun Grannis. Health-Terrain: A Visual Analytics System for Health Data. Journal of American Medical Informatics Association, Accepted.

Shenghui Jiang, Shiaofen Fang, Sam Bloomquist, Jeremy Keiper, Mathew Palakal, Yuni Xia, and Shaun Grannis. Healthcare Data Visualization: Geospatial and Temporal Integration. To Appear: Proc. of 2016 International Conference on Information Visualization, Theory and Applications (IVAPP), 2016

J. Keiper, Shiaofen Fang, Yuni. Xia, Mathew Palakal, R. Shaun Grannis, Thanh Minh Nguyen, Sam Bloomquist, Anand Krishnan, Weizhi Li. A Public Health Data Visualization System Demonstration, AMIA 2014, Demo paper, Washington DC. Nov. 2014.

Y. Xia, S. Fang, M. Palakal, R. Gamache, T. Nguyen, S. Bloomquist, J. Keiper, S. Grannis. Data Exploration of a Notifiable Condition Detector System. In Proc. 2013 Workshop on Visual Analytics in Healthcare. Poster Presentation, Washington DC, Oct. 2013, pp 66-67.

M. Palakal, S. Fang, Y. Xia, S. Grannis, R. Gamache, T. Nguyen, S. Bloomquist, J. Keiper. Detecting Comorbidity of Chlamydia from Clinical Reports. In Proc. 2013 Workshop on Visual Analytics in Healthcare. Poster Presentation, Washington DC. Oct. 2013, pp 75-76.

Jeremy Keiper, Yuni Xia, Shiaofen Fang,et al. Use Cases for Public Health Data Visualization. In Proc. 2013 Workshop on Visual Analytics in Healthcare. Washington DC. Oct. 2013.

Presentations

J. Keiper, Shiaofen Fang, Yuni. Xia, Mathew Palakal, R. Shaun Grannis, Thanh Minh Nguyen, Sam Bloomquist, Anand Krishnan, Weizhi Li. A Public Health Data Visualization System Demonstration, AMIA 2014, Demo paper, Washington DC. Nov. 2014.

J. Keiper, Y. Xia, S. Fang, M. Palakal, R. Gamache, T. Nguyen, S. Bloomquist, J. Keiper, S. Grannis. Use Cases for Public Health Data Visualization. In Proc. 2013 Workshop on Visual Analytics in Healthcare. Poster Presentation, Washington DC. Oct. 2013.

Y. Xia, S. Fang, M. Palakal, R. Gamache, T. Nguyen, S. Bloomquist, J. Keiper, S. Grannis. Data Exploration of a Notifiable Condition Detector System. In Proc. 2013 Workshop on Visual Analytics in Healthcare. Poster Presentation, Washington DC, Oct. 2013.

M. Palakal, S. Fang, Y. Xia, S. Grannis, R. Gamache, T. Nguyen, S. Bloomquist, J. Keiper. Detecting Comorbidity of Chlamydia from Clinical Reports. In Proc. 2013 Workshop on Visual Analytics in Healthcare. Poster Presentation, Washington DC. Oct. 2013.

INVENTIONS, PATENTS AND LICENSES

Nothing to report.

REPORTABLE OUTCOMES

1. We have constructed 5 ontologies, one for each category - Disease, Symptom, Mental behavior, Risky Behavior and Medication based on the data extracted from the clinical notes. This helps us in eliminating noise and providing structure to our findings.
2. We have built a specialized database for the storage and real time query of the concept space and the NCD dataset. The database is general enough to be adapted to any other health care data and extensions of the concept space.
3. We have developed a prototype visualization system for healthcare data. The system provides a web based user interface that allows interactive visualization and exploration of a given dataset representing a use case.
4. We published 6 research papers and made 4 presentations based on the work in this project.

OTHER ACHIEVEMENTS

Nothing to report

REFERENCES

1. Automated Electronic Lab Reporting and Case Notification, last retrieved from http://www.regenstrief.org/cbmi/areas-excellence/public-health/
2. Fighting disease outbreaks with two-way health information exchange, last retrieved from http://newsinfo.iu.edu/news/page/normal/11948.html
3. B.L. Humphreys, D.A. Lindberg, H.M. Schoolman, G.O. Barnett. The unified medical language system: An informatics research collaboration J. Am. Med. Inform. Assoc., 5 (1) (1998), pp. 1–11
4. Chapman , W.; Bridewell , ; Hanbury , ; Cooper , G. F.; Buchanan , G. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics 2001, 34 (5), 301–310.
5. Palakal M., Stephens M., Mukhopadyay S., Raje R. Identification of biological relationships from text documents using efficient computational methods, Journal of Bioinformatics and Computational Biology, Vol. 1, No. 2(2003) 307-342
6. Van Mechelen I, Bock HH, De Boeck P. Two-mode clustering methods:a structured overview. Statistical Methods in Medical Research 13 (5): 363–94, 2004
7. [Stephen G. Kobourov. Spring Embedders and Force Directed Graph Drawing Algorithms. arXiv: 1201.3011.
8. S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "ThemeRiver: Visualizing Thematic Changes in Large Document Collections," Visualization and Computer Graphics, IEEE Transactions, vol. 8, pp. 9-20, 2002
9. Weber, M., Marc Alexa, Wolfgang Müller. Visualizing Time-Series on Spirals. IEEE Information Visualization, 7-13, 2001.
10. Nathan Gossett, Baoquan Chen. Paint Inspired Color Mixing and Compositing for Visualization. IEEE Symposium on Information Visualization 2004. 113-117
11. Ken Perlin. An image synthesizer. In Proceedings of SIGGRAPH85, pages 287–296. ACM Press, 1985.
12. Rosenfeld, A. and A.C. Kak (1982). Digital Picture Processing. Academic Press, New York.
13. Hoschek, J., (1988), “Spline Approximation of Offset Curves,” Computer Aided Geometric Design, Vol. 5, pp. 33–40.
14. You, Q., Fang, S., Chen, J. GeneTerrain: Visual Exploration of Differential Gene Expression Profiles Organized in Native Biomolecular Interaction Networks. Journal of Information Visualization, 2010; 9:1, 1-12.

APPENDICES

Attach all appendices that contain information that supplements, clarifies or supports the text. Examples include original copies of journal articles, reprints of manuscripts and abstracts, a curriculum vitae, patent applications, study questionnaires, and surveys, etc.